

Wi-Fi LIVE ONLINE TRAINING

Web Scraping in 60 Minutes



MAX HUMBER

<https://resources.oreilly.com/binderhub/web-scraping-in-60-minutes>

November 3, 2020

12:00pm – 1:00pm EDT

1

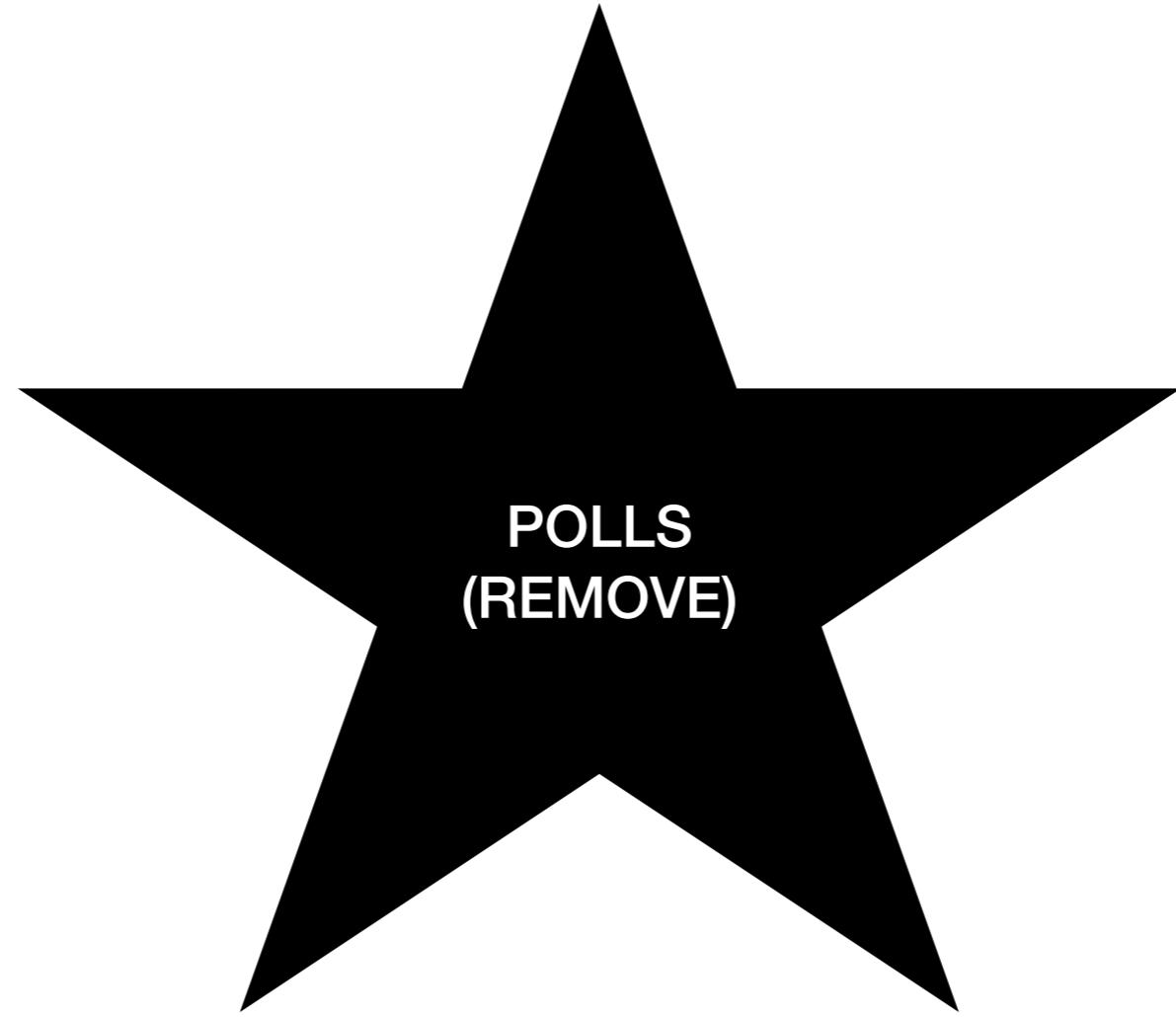


2

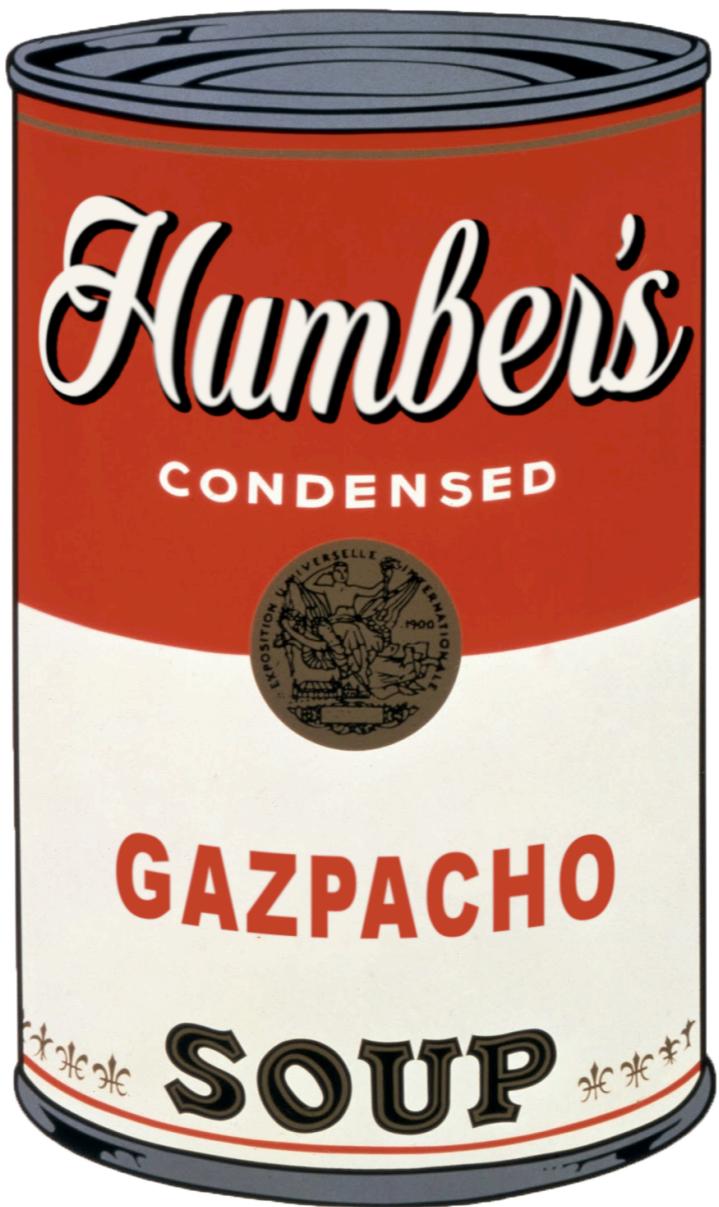


3





- Poll: How many websites have you scraped before? {0, 1, 10, 100+}
- Poll: Is your interest in web scraping professional or personal? {professional, personal}



<https://gazpacho.xyz>



33



5



57

...



BeautifulSoup is so 2000-and-late: Web Scraping in 2020

#python

#webscraping

#gazpacho

#hacktoberfest

- 
1. No dependencies
 - 2. Batteries included**
 - 3. 1 find method**
 4. Production ready
 5. PEP 561 compliant
 6. Automatic formatting
 - 7. Speed (50% faster)**
 - 8. Partial matching**
 9. Debt-free (3 first)
 10. Open (and Friendly)

> 500/week
> 450 now



Max Humber

Human

1mo •

...

🎉 gazpacho JUST TURNED 1 🎉

Some highlights from this year:

20,000 installs

300 stars

Used by teams at Apple, Microsoft, and Facebook

To mark the occasion I've just bumped the library out of ZeroVer hell and released version 1.0!

Improvements include:

type: hints = everywhere

The ability to parse malformed void tags

A new way to initialize a parser with: Soup.get("url")

Details: <https://gazpacho.xyz/>

And if you'd like to help build out the roadmap for gazpacho please fill out this quick (seven questions, all optional) survey:

<https://lnkd.in/gKWzRqM>



Max Humber

Human

2w •

...

gazpacho is #1 on **#Hacktoberfest** right now

After a full year of development it finally feels like the *most beautiful soup* 😊 has some real momentum behind it

The screenshot shows a web browser window with the title bar "Hacktoberfest presented by DigitalOcean". The address bar displays the URL "https://hacktoberfest.digitalocean.com". The main content area is titled "Hacktoberfest projects" and contains a message: "You can contribute to participating projects on GitHub. Here are a few looking for some help:". Below this, there is a dropdown menu labeled "Select language". The page then lists several GitHub repositories:

| Project Name | Language | Description |
|--------------------------|------------|---|
| gazpacho | Python | The simple, fast, and modern web scraping library |
| python-backgroundchanger | Python | A simple background changer using Unsplash API |
| LeetCode_Algorithms | Python | A collection of solutions for Medium/Hard LeetCode problems. Educational resource |
| jina | Python | An easier way to build neural search in the cloud |
| tmdb-app | JavaScript | An unofficial client app to The Movie Database made with React Native |
| github-readme-mediumizer | JavaScript | Post medium card in Github README |

Each project card includes a small image, the repository name, its language, and a brief description. Some cards also show pull request counts and descriptions.



Q&A

That's all Folks!

gazpacho

Formats ▾

Publication Date ▾

Topics ▾

Publishers ▾

Rating ▾

Sort By Relevance ▾

1 - 10 of 39 search results for gazpacho



</> SCENARIO

Using gazpacho to Download and Parse the Contents of a Website

By [Max Humber](#)

WEB PERFORMANCE

O'Reilly Media, Inc. June 2020

Scrape a simple web page with gazpacho



</> SCENARIO

Using gazpacho and pandas to Retrieve the Contents of an HTML Table

By [Max Humber](#)

PANDAS

O'Reilly Media, Inc. June 2020

Scrape an HTML table with pandas

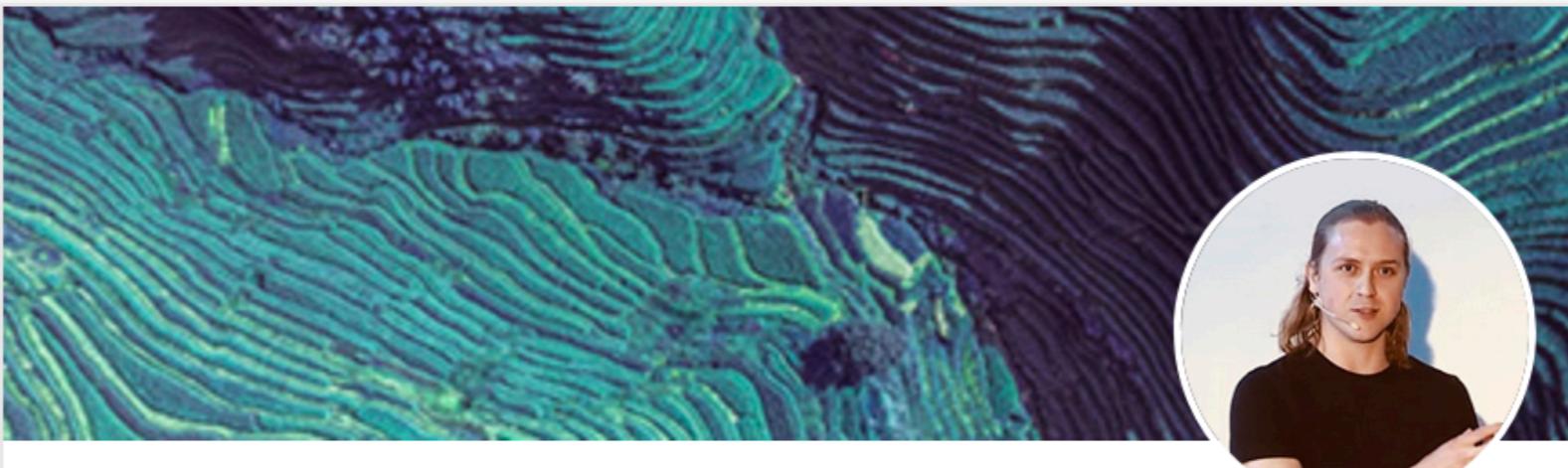


</> SCENARIO

Using gazpacho and Selenium to Retrieve the Contents of a Password-Protected Web Page

By [Max Humber](#)

SELENIUM



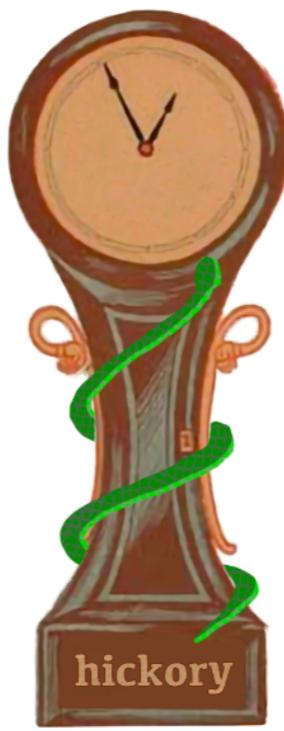
November 10, 2020

Advanced Web Scraping

Presented by Max Humber

78 SPOTS REMAINING

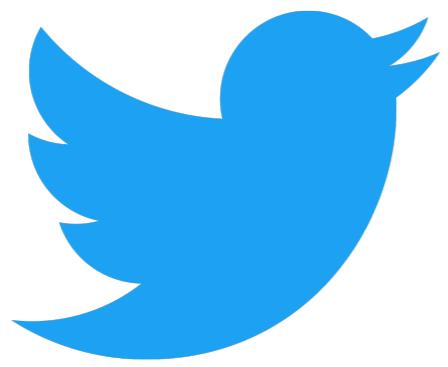
LEARN MORE



github.com/maxhumber/hickory



github.com/maxhumber/gif



twitter.com/maxhumber



www.linkedin.com/in/maxhumber